AD-A080 715 WISCONSIN UNIV-MADISON DEPT OF STATISTICS F/G 12/1
TRANSFORMING GROUPED BIVARIATE DATA TO NEAR NORMALITY.(U)
NOV 79 V M GUERRERO, R A JOHNSON N00014-78-C-0722
UNCLASSIFIED TR-592 ARO-15660.3-M NL

1 OF 1
AD
A080715

END
DATE
FILMED
3 - 80
DDC

18 ARO 15660.3-M

19

DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN

Madison, Wisconsin

LEVEL II

14 TR-51

TECHNICAL REP[redacted]T NO. 592

12 22

Nov[redacted]79

TRANSFORMING GROUPED BIVARIATE DATA TO NEAR NORMALITY.

by

Victor M. Guerrero [redacted] Richard A. Johnson
University of Wisconsin

15 N00014-78-C-0722

TRANSFORMING GROUPED BIVARIATE DATA TO NEAR NORMALITY

by

Victor M. Guerrero* and Richard A. Johnson

University of Wisconsin

Key Words: grouped bivariate data
transformations
regression and correlation

This research was sponsored by the Office of Naval Research under Grant No. N00014-78-C-0722. (Also funded by Army Research Office)

*Currently at Direccion General de Programacion - SEP. Mexico

Transforming Grouped Bivariate Data to Near Normality

by

Victor M. Guerrero and Richard A. Johnson

Abstract

We are concerned with the use of power transformations when data on two variables are presented in a two way table. Regressions and the correlation are obtained from the transformed grouped data. Also, by transforming back to the original scale, we obtain a smoothed version of the data.

1. Introduction

Reexpression of data through transformations can enhance understanding. We propose methods for selecting a transformation for bivariate data that are grouped. Grouping commonly arises in census or surveys where responses consist of checking appropriate intervals. Other data are sometimes grouped to avoid biases like that caused by people reporting their age to the nearest five years. Transformations to normality should improve the degree of association as well as improve marginal normality.

When two continuous random variables are classified as a two-way table, these tables are called correlation tables (c.f. Kendall and Buckland (1960)). The most successful approaches for expressing association are those which assume a parametric form for an underlying distribution. Because the assumed distribution is often bivariate normal, we investigate transformations to near normality. In section 2, we develop a technique for interpreting association as the correlation determined from grouped data that are first transformed to near normality.

The related bivariate normal regression problem, discussed in Section 3, is concerned with obtaining a "good" prediction equation. In Section 4, we study simple linear regression, where one of the variables can be treated as fixed. Examples illustrate each technique .

2. Correlation from Transformed Grouped Data

2.1 Procedure for Selecting a Transformation

A transformation to near bivariate normality should aid in the interpretation of association. Lancaster (1957) established that if (X,Y) are jointly normal, then the correlation of any transformed variables $X' = X'(X)$ and $Y' = Y'(Y)$ is smaller in absolute value than that of X and Y. It is hoped that transforming grouped data to near normality will also produce a nearly largest correlation.

Consider a sample of n independent pairs $(X_{11},X_{21}),\dots,(X_{1n},X_{2n})$ from an absolutely continuous distribution with pdf g concentrated on $(0,\infty)\times(0,\infty)$. As with the usual Box and Cox (1964) approach, let us suppose (pretend) that a vector parameter $\underset{\sim}{\theta}_0 = (\mu_{10},\sigma_{10},\mu_{20},\sigma_{20},\rho_0,\lambda_{10},\lambda_{20})'$ exists such that

$$\underset{\sim}{X}^{(\underset{\sim}{\lambda}_0)} = \begin{pmatrix} X_1^{(\lambda_{10})} \\ X_2^{(\lambda_{20})} \end{pmatrix} \sim N_2(\underset{\sim}{\mu}_0,\Sigma_0)$$

where

$$X_i^{(\lambda_{i0})} = \begin{cases} \dfrac{X_i^{\lambda_{i0}} - 1}{\lambda_i} & \text{if } \lambda_{i0} \neq 0 \\ \log(X_i) & \text{if } \lambda_{i0} = 0 \end{cases} \qquad i = 1,2 .$$

Let the sample be grouped into k·h cells, specified beforehand and determined by $k \geq 3$ intervals (with endpoints $0 = a_0 < a_1 < \dots < a_{k-1} < a_k = \infty$) on the x_1 axis, and by $h \geq 3$ intervals (with endpoints $0 = b_0 < b_1 < \dots < b_{h-1} < b_h = \infty$) on the x_2 axis. The notational convention $a_0^{(\lambda)} = b_0^{(\lambda)} = -\infty$, for all values of λ, is used to simplify some expressions.

If we denote the number of observations falling into the i-j^{th} cell by n_{ij}, then we can write the likelihood function for a sample of size $n = \sum_{k=1}^{k}\sum_{j=1}^{h} n_{ij}$, as

$$L_n(\underset{\sim}{\theta}) = (n!/\prod_{i=1}^{k}\prod_{j=1}^{h} n_{ij}!)\prod_{i=1}^{k}\prod_{j=1}^{h} p_{ij}^{n_{ij}}(\underset{\sim}{\theta}) \tag{2}$$

where

$$p_{ij}(\underset{\sim}{\theta}) = \Phi_2\left(\frac{a_i^{(\lambda_1)} - \mu_1}{\sigma_1}, \frac{b_j^{(\lambda_2)} - \mu_2}{\sigma_2}\right) + \Phi_2\left(\frac{a_{i-1}^{(\lambda_1)} - \mu_1}{\sigma_1}, \frac{b_{j-1}^{(\lambda_2)} - \mu_2}{\sigma_2}\right)$$
$$- \Phi_2\left(\frac{a_i^{(\lambda_1)} - \mu_1}{\sigma_1}, \frac{b_{j-1}^{(\lambda_2)} - \mu_2}{\sigma_2}\right) - \Phi_2\left(\frac{a_{i-1}^{(\lambda_1)} - \mu_1}{\sigma_1}, \frac{b_j^{(\lambda_2)} - \mu_2}{\sigma_2}\right) \tag{3}$$

and Φ_2 is the bivariate normal cdf with zero means, unit variances and correlation ρ.

The large sample properties of the maximum likelihood estimator (MLE) $\hat{\underset{\sim}{\theta}}_n$ are given by Theorem 1 below. Its proof consists of first establishing the consistency of $\hat{\underset{\sim}{\theta}}_n$ by showing that the log-likelihood converges uniformly to a limit. This log-likelihood has sufficiently smooth derivatives so that asymptotic normality is obtained from a Taylor expansion.

<u>**Theorem 1.**</u> Let $q_{ij} = \int_{a_{i-1}}^{a_i} \int_{b_{j-1}}^{b_j} g(x_1,x_2)dx_2dx_1$ for all i,k. Let $p_{ij}(\underset{\sim}{\theta})$ be given by (3) and assume that

(i) the parameter space Ω is a compact subset of $\mathbb{R}^7$,

(ii) $H(\underset{\sim}{\theta}) = \sum_{i=1}^{k} \sum_{j=1}^{h} q_{ij} \log \left[\frac{p_{ij}(\underset{\sim}{\theta})}{q_{ij}}\right]$ has a unique global maximum as a function of $\underset{\sim}{\theta} = (\mu_1,\sigma_1,\mu_2,\sigma_2,\rho,\lambda_1,\lambda_2)'$ and this is attained at $\underset{\sim}{\theta} = \underset{\sim}{\theta}_0$.

Then, (1) $\hat{\underset{\sim}{\theta}}_n \xrightarrow{a.s.} \underset{\sim}{\theta}_0$ as $n \to \infty$.

Furthermore, if

(iii) $\underset{\sim}{\theta}_0$ is an interior point of Ω

(iv) the Hessian of $H(\underset{\sim}{\theta})$ is nonsingular at $\underset{\sim}{\theta}_0$.

Then, (2) $\sqrt{n}(\hat{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta}_0) \xrightarrow{d} N_7(\underset{\sim}{0}, VWV')$ as $n \to \infty$ where

$$V = [\nabla^2 H(\underset{\sim}{\theta}_0)]^{-1} \quad \text{and}$$

$$W = \left(\sum_{i=1}^{k} \sum_{j=1}^{h} q_{ij} \left(\frac{\partial \log[p_{ij}(\underset{\sim}{\theta})]}{\partial \theta_u}\right)\left(\frac{\partial \log[p_{ij}(\underset{\sim}{\theta})]}{\partial \theta_v}\right)\right)_{7\times 7}.$$

□

2.2 Iterative Procedure and Likelihood Equations

To apply our procedure we suggest using the two-stage method of Richards (1961) in order to obtain the MLE $\hat{\underset{\sim}{\theta}}_n$. Explicitly, we propose to maximize the log-likelihood $\ell_n(\underset{\sim}{\theta})$ as follows: first, fix λ_1 and λ_2 and maximize over the values of the remaining parameters; then search for a global maximum of $\ell_n(\underset{\sim}{\theta})$ on a grid of values of λ_1 and λ_2.

In order to present a numerical example, we will first find the maximum likelihood equations for the parameters μ_1, σ_1, μ_2, σ_2 and ρ.

Set $A_i = (a_i^{(\lambda_1)} - \mu_1)/\sigma_1$, $i = 0,1,\ldots,k$ and $B_j = (b_j^{(\lambda_2)} - \mu_2)/\sigma_2$, $j = 0,\ldots,h$. Expressing the joint pdf as the product of marginal and conditional pdf's, it is easily show that

$$\frac{\partial \ln L_n}{\partial \mu_1} = \sum_{i=1}^{k} \sum_{j=1}^{h} n_{ij}[\sigma_1 p_{ij}(\underset{\sim}{\theta})]^{-1}(\phi(A_i)[\Phi(F_{j-1,i}) - \Phi(F_{j,i})]$$
$$- \phi(A_{i-1})[\Phi(F_{j-1,i-1}) - \Phi(F_{j,i-1})])$$

$$\frac{\partial \ln L_n}{\partial \sigma_1} = \sum_{i=1}^{k} \sum_{j=1}^{h} n_{ij}[\sigma_1 p_{ij}(\underset{\sim}{\theta})]^{-1} (A_i\phi(A_i)[\Phi(F_{j-1,i}) - \Phi(F_{j,i})]$$
$$- A_{i-1}\phi(A_{i-1})[\Phi(F_{j-1,i}) - \Phi(F_{j,i-1})])$$

$$\frac{\partial \ln L_n}{\partial \rho} = \sum_{i=1}^{k} \sum_{j=1}^{h} n_{ij}\, p_{ij}^{-1}(\underset{\sim}{\theta})(\phi_2(A_i,B_j) + \phi_2(A_{i-1},B_{j-1}) - \phi_2(A_i,B_{j-1})$$
$$- \phi_2(A_{i-1},B_j))$$

where $F_{j,i} = (B_j - \rho A_i)(1-\rho^2)^{-1/2}$ for each i and j and similar expressions hold for $\partial \ln L_n/\partial \mu_2$ and $\partial \ln L_n/\partial \sigma_2$. These equations are to be solved iteratively for fixed values of (λ_1,λ_2).

2.3 Example

The data used for illustrative purposes are the "Age of Parents of Boys" given by Cramer (1947, p. 458.) Our procedure for finding the "closest" bivariate normal distribution produced the transformation parameters $(\hat{\lambda}_1,\hat{\lambda}_2) = (-.04, .42)$ and the parameter estimates $(\mu_1,\sigma_1,\mu_2,\sigma_2,\rho) = (3.3068, .2130, 7.7940, .9384, .6810)$ for the approximating normal. Initial estimates for the parameters $(\lambda_1,\mu_1,\sigma_1)$ and $(\lambda_2,\mu_2,\sigma_2)$ were obtained by first transforming

each marginal distribution to near normality. Cramer (1947) originally noticed that the marginal distributions were obviously non-normal and suggested taking the log of the father's age. Our results corroborate this conjecture.

Contours of the bivariate normal density are shown in Figure 1 together with the relative cell frequencies. Figure 2 presents contours of the "inverted" bivariate normal and the corresponding relative cell frequencies.

A value of $\chi^2 = 46{,}034$ rejects the adequacy of the fitted normal. However, the transformation does greatly improve the normal approximation.

3. Bivariate Regression from Transformed Grouped Data

Let us consider the case of estimating the regression of X_1 on X_2, as the other case follows by symmetry. Under the tentative assumption (1), we have

$$\mathcal{L}(X_1^{(\lambda_1)} \mid X_2^{(\lambda_2)} = x_2^{(\lambda_2)}) = N(\mu_1 + \beta_{12}(x_2^{(\lambda_2)} - \mu_2),\ \sigma_1^2(1-\rho^2))$$

so that

$$E(X_1^{(\lambda_1)} \mid X_2^{(\lambda_2)} = x_2^{(\lambda_2)}) = \mu_1 + \beta_{12}(x_2^{(\lambda_2)} - \mu_2). \tag{5}$$

Also $X_2^{(\lambda_2)}$ is $N(\mu_2, \sigma_2^2)$. By the invariance of MLE's we obtain the MLE $\hat{\underline{\theta}}_n^* = (\hat{\mu}_{1n}, \hat{\sigma}_{1n}, \hat{\rho}_n \frac{\hat{\sigma}_{1n}}{\hat{\sigma}_{2n}}, \hat{\mu}_{2n}, \hat{\sigma}_{2n}, \hat{\lambda}_{1n}, \hat{\lambda}_{2n})'$ for $\theta^* = (\mu_1, \sigma_1, \beta_{12}, \mu_2, \sigma_2, \lambda_1, \lambda_2)'$.

From the asymptotic properties of $\hat{\underline{\theta}}_n$ we easily conclude $\lim_{n\to\infty} \hat{\underline{\theta}}_n^* = \underline{\theta}_0^*$ with probability one and

$$\sqrt{n}(\hat{\underline{\theta}}_n^* - \underline{\theta}_0^*) \xrightarrow{d} N_7(\underline{0},\ RVWV'R') \text{ as } n \to \infty$$

where $R = (\partial\theta_i^*/\partial\theta_j)_{7\times 7}$ and VWV' is the matrix appearing in Theorem 1.

The following numerical illustration uses the data on "Age of Parents of Boys" introduced previously. The parameter estimates for the regression of

X_1 on X_2 become $\hat{\mu}_1 = 3.3068$, $\hat{\sigma}_1 = .2130$, $\hat{\beta}_{12} = \hat{\rho}\, \hat{\sigma}_1/\hat{\sigma}_2 = .1546$, $\hat{\mu}_2 = 7.7940$, $\hat{\sigma}_2 = .9384$, $\hat{\lambda}_1 = -.04$ and $\hat{\lambda}_2 = .42$. For the regression of X_2 on X_1 we obtain $\hat{\beta}_{21} = \hat{\rho}\, \hat{\sigma}_2/\hat{\sigma}_1 = 2.9999$. The normal regression lines, obtained in the transformed scale, appear in Figure 3. The corresponding regression curves, after transforming back to the original scale, are shown in Figure 4.

4. Simple Linear Regression

We now investigate *fitting a linear model to a (transformed) pair of* variables $Y^{(\lambda_1)}$ and $x^{(\lambda_2)}$, where one or both variables are presented in grouped form, but x is treated as fixed. In the ungrouped-data situation, the data are a sample of n pairs $(y_1, x_1), \ldots, (y_n, x_n)$ where the y's are realizations of the random variable Y and the x's are some specified fixed values at which Y was observed. The simple linear regression model has the form

$$\begin{aligned} Y_u^{(\lambda_1)} &= \alpha + \beta x_u^{(\lambda_2)} + \varepsilon_u, \qquad u = 1, \ldots, n \\ \varepsilon_u &\sim N(0, \sigma^2), \quad \mathrm{Cov}(\varepsilon_u, \varepsilon_v) = 0 \quad \text{if } u \neq v \end{aligned} \tag{6}$$

where α, β and σ are the regression parameters which must be estimated in conjunction with the transformation parameter λ_1 and λ_2.

The model (6) provides a good fit to some economic data. For instance for family budgets, an investigator wishes to determine the relationship between the expenditure (y) on a particular commodity and the income level. of the household (x). This relationship, termed the Engel curve, (c.f. Prais and Houthakker (1955)) *has been modelled as* $y = \alpha + \beta x$, $y = \alpha + \beta \log(x)$, $y = \alpha + \beta/x$, $\log(y) = \alpha + \beta \log(x)$, or $\log(y) = \alpha + \beta/x$.

Originally the choice of a particular functional form was often made on the

basis of more or less ad hoc tests, but Benus, Kmenta and Shapiro (1976) have shown how to use the Box-Cox method to obtain a more objective choice.

In our grouped-data case, we assume the existence of a parameter vector $\underset{\sim}{\theta} = (\alpha,\beta,\sigma,\lambda_1,\lambda_2)'$ for which (6) holds. Two grouping schemes, that have been studied by other authors (c.f. Fryer and Pethybridge (1972)),

(1) Both Variables in Grouped Form

(2) Only One Variable Grouped

are considered.

4.1 Both Variables in Grouped Form

Suppose that the data are given in a frequency table like

y \ x	$[a_0,a_1)$	$[a_1,a_2)$	. . .	$[a_{k-1},a_k)$	Total
$[b_0,b_1)$	n_{11}	n_{12}	. . .	n_{1k}	$n_{1\cdot}$
$[b_1,b_2)$	n_{21}	n_{22}	. . .	n_{2k}	$n_{2\cdot}$
. . .	. . .	. . .	. . .	. . .	. . .
$[b_{h-1},b_h)$	n_{h1}	n_{h2}	. . .	n_{hk}	$n_{h\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	. . .	$n_{\cdot k}$	n

where $a_0 = b_0 = 0$ and $a_k = b_h = \infty$. For a given sample size n, the quantities $n_{\cdot i}$ are supposed to be fixed. In order to carry out calculations with grouped data, it is customary to assume that the data are placed at the midpoints of the intervals. For the independent variable $x_u \in [a_{i-1},a_i)$ we use the midpoint $z_i(\lambda_2) = \dfrac{a_i^{(\lambda_2)} + a_{i-1}^{(\lambda_2)}}{2}$ instead of $x_u^{(\lambda_2)}$. Model (6) suggests

$$Y_u^{(\lambda_1)} \sim N(\alpha+\beta\, z_i(\lambda_2),\sigma^2) \tag{7}$$

for each u such that $x_u \in [a_{i-1},a_i)$, $i = 1,\ldots,k$ and $u = 1,\ldots,n$.

Since Y_u is independent of Y_v for $u \neq v$, the log-likelihood of the sample becomes

$$\ell_n(\underset{\sim}{\theta}) = \sum_{i=1}^{k} [\log(n_{\cdot i}!) - \sum_{j=1}^{h} \log(n_{ji}!)] + \sum_{i=1}^{k}\sum_{j=1}^{h} n_{ji}\log[p_{ji}(\underset{\sim}{\theta})].$$

where

$$p_{ji}(\underset{\sim}{\theta}) = \Phi\left(\frac{b_j^{(\lambda_1)} - \alpha - \beta\, z_i(\lambda_2)}{\sigma}\right) - \Phi\left(\frac{b_{j-1}^{(\lambda_1)} - \alpha - \beta\, z_i(\lambda_2)}{\sigma}\right). \tag{8}$$

In order to adequately model a regression situation, we assume that the true underlying distribution of Y varies according to the levels of x. That is, we assume Y_u has pdf g_i (concentrated on $(0,\infty)$) whenever x_u lies in $[a_{i-1},a_i)$. Then the true probabilities for the intervals $[b_{j-1},b_j)$ become

$$q_{ij} = \int_{b_{j-1}}^{b_j} g_i(y)dy \quad \text{for } i = 1,\ldots,k \text{ and } j = 1,\ldots,h. \tag{9}$$

We then have

<u>Theorem 2.</u> Let $p_{ji}(\underset{\sim}{\theta})$ and q_{ji} be as in (8) and (9). Let $r_1,\ldots,r_k$ be some fixed numbers such that $0 < r_i < 1$ $\forall i$ and $\sum_{i=1}^{k} r_i = 1$, and suppose

(i) the parameter space $\Omega \subset \mathbb{R}^5$ is compact

(ii) $\lim_{n\to\infty} \frac{n_{\cdot i}}{n} = r_i$ for $i = 1,\ldots,k$

(iii) $H(\underset{\sim}{\theta}) = \sum_{i=1}^{k}\sum_{j=1}^{h} r_i q_{ji} \log\left[\frac{p_{ji}(\underset{\sim}{\theta})}{q_{ji}}\right]$ attains a unique global maximum at $\underset{\sim}{\theta}_0 = (\alpha_0,\beta_0,\sigma_0,\lambda_{10},\lambda_{20})'$.

Then, (1) $\hat{\underset{\sim}{\theta}}_n \xrightarrow{a.s.} \underset{\sim}{\theta}_0$ as $n \to \infty$.

Moreover, if

(iv) $\underset{\sim}{\theta}_0$ is an interior point of Ω

(v) the Hessian of $H(\underset{\sim}{\theta})$ is nonsingular at $\underset{\sim}{\theta}_0$.

Then, (2) $\sqrt{n}(\hat{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta}_0) \xrightarrow{\mathcal{D}} N_5(\underset{\sim}{0}, VWV')$ as $n \to \infty$, where

$$V = [\nabla^2 H(\underset{\sim}{\theta})]^{-1} \text{ and}$$

$$W = \left(\sum_{i=1}^{k}\sum_{j=1}^{h} r_i\, q_{ji} \left(\frac{\partial \log[p_{ji}(\underset{\sim}{\theta})]}{\partial\theta_u}\right)\left(\frac{\partial \log[p_{ji}(\underset{\sim}{\theta})]}{\partial\theta_v}\right)\right)_{5\times 5}.$$

Proof: See Guerrero (1979)

4.2 Only One Variable Grouped

Let us consider first the case in which only the dependent variable is grouped. Suppose that $n_{\cdot i}$ observations of Y are made at the fixed value x_i, where $i = 1,\ldots,k$ and $\sum_{i=1}^{k} n_{\cdot i} = n$ is the total sample size. This situation is like the previous case, but with the (approximate) quantities $z_i(\lambda_2)$ replaced by the (exact) transformed observation $x_i^{(\lambda_2)}$.

The other case is when the outcomes of the random variable Y are exactly specified, but the values of x are grouped into the intervals $[0,a_1),[a_1,a_2),\ldots,[a_{k-1},\infty)$. In the latter case we again consider the model (7) which gives rise to the log-likelihood

$$\begin{aligned}\ell_n(\underset{\sim}{\theta}|\underset{\sim}{y}) = &-\frac{n}{2}\log(2\pi) - n\log(\sigma) \\ &- \frac{1}{2\sigma^2}\sum_{u=1}^{n}\sum_{i=1}^{k}\delta_{iu}(y_u^{(\lambda_1)} - \alpha - \beta\, z_i(\lambda_2))^2 \\ &+ n(\lambda_1 - 1)\log(y_u).\end{aligned} \tag{10}$$

where $\delta_{iu} = 1$ if $x_u \in [a_{i-1}, a_i)$ and 0 otherwise.

Now, letting $n_i = \sum_{u=1}^{n}\delta_{iu}$, $i = 1,\ldots,k$ and assuming that Y_u has pdf g_i whenever x_u belongs to $[a_{i-1}, a_i)$, we are able to establish the following.

Theorem 3. Let $r_1,\ldots,r_k$ be some fixed numbers such that $0 < r_i < 1$ $\forall i$ and $\sum_{i=1}^{k} r_i = 1$. If

(i) the parameter space Ω is the compact set given by

$\Omega = \{\underset{\sim}{\theta} = (\alpha,\beta,\sigma,\lambda_1,\lambda_2)' \,|\, |\alpha| \le M_1,\ |\beta| \le M_2,\ s_1 \le \sigma \le s_2,$
$a \le \lambda_1 \le b,\ c \le \lambda_2 \le d$, with $0 < M_1, M_2, s_1, s_2, b, d < \infty$ and $-\infty < a, c < 0\}$

(ii) $E_{g_i}(Y^{2a})$ and $E_{g_i}(Y^{2b})$ are both finite $\forall i$

(iii) $\lim_{n\to\infty} \frac{n_i}{n} = r_i$, $i = 1,\ldots,k$

(iv) $\sum_{i=1}^{k} r_i\, E_{g_i}[\ell_1(\underset{\sim}{\theta}|Y)]$ has a unique global maximum at $\underset{\sim}{\theta} = \underset{\sim}{\theta}_0$.

Then (1) $\hat{\underset{\sim}{\theta}}_n \xrightarrow{a.s.} \underset{\sim}{\theta}_0$ as $n \to \infty$.

Furthermore, if

(v) $\underset{\sim}{\theta}_0$ is an interior point of Ω

(vi) both $E_{g_i}[Y^a \log(Y)]^2$ and $E_{g_i}[Y^b \log(Y)]^2$ are finite $\forall i$

(vii) $\sum_{i=1}^{k}\sqrt{r_i}\; E_{g_i}[\nabla\ell_1(\underset{\sim}{\theta}_0|Y)] = \underset{\sim}{0}$

(viii) $V = \{\sum_{i=1}^{k} r_i\, E_{g_i}[\nabla^2 \ell_1(\underset{\sim}{\theta}_0|Y)]\}^{-1}$ exists.

Then (2) $\sqrt{n}(\hat{\underline{\theta}}_n - \underline{\theta}_0) \xrightarrow{d} N_5(\underline{0}, VWV')$ where

$$W = \sum_{i=1}^{k} r_i \, \mathrm{Var}_{g_i}[\nabla \ell_i(\underline{\theta}_0|Y)].$$

Proof: See Guerrero (1979).

When Y is grouped (and x is either grouped or ungrouped) we do not require the specific functional forms of the g_i's. Moreover, the true probabilities g_{ij}, defined in (9) are consistently estimated by the observed frequencies $n_{ij}/n_{\cdot i}$, $\forall i,j$.

4.3 Numerical Example

The following illustration shows how to fit a simple linear regression model when both y and x are given in grouped form. We again employ a two-stage maximization procedure to obtain the global maximum of the log-likelihood. For more information on the computational aspects, see Guerrero (1979).

Table 1 gives the observed frequencies in several groups of total miles driven (y) and family income (x) during 1973 (for car owners). If miles driven are considered as indirect observations of amount of money spent on gasoline, then the problem of fitting a regression line to the transformed variables $y^{(\lambda_1)}$ and $x^{(\lambda_2)}$ becomes that of fitting an Engel function to y and x.

The MLE's are $\hat{\lambda}_1 = .375$, $\hat{\lambda}_2 = .356$, $\hat{\alpha} = .50327$, $\hat{\beta} = .05135$ and $\hat{\sigma} = 1.9004$.

Figure 5 shows the fitted regression line in the transformed scale. Transforming back to the original scale, we obtain the curve shown in Figure 6.

Bibliography

Benus, J., Kmenta, J. and Shapiro, H. (1976). "The dynamics of household budget allocation to food expenditures." Rev. of Econ. and Statist. 58, 129-38.

Box, G.E.P and Cox, D.R. (1964). "An analysis of transformations." J.R. Statist . Soc. B-26, 211-52.

Cramér, H. (1947). Mathematical Methods of Statistics. Princeton University Press.

Fryer, J.G. and Pethybridge, R.J. (1972). "Maximum likelihood estimation of a linear regression function with grouped data." Appl. Statist. 21, 142-54.

Guerrero, G.V.M. (1979). "Extensions of the Box-Cox Transformation to *Grouped Data Situations*". Ph.D. Thesis, University of Wisconsin-Madison.

Holmes, J. (1974). "The relative burden of higher gasoline prices." In Five Thousand American Families-Patterns of Economic Progress. Vol. IV. Ann Arbor, Institute for Social Research.

Kendall, M.G. and Buckland, W.R. (1960). A Dictionary of Statistical Terms. Hafner Publishing Co.

Lancaster, H.O. (1957). "Some properties of the bivariate normal distribution considered in the form of a contingency table." Biometrika 44, 282-92.

Prais, S.J. and Houthakker, H.S. (1955). The Analysis of Family Budgets. Cambridge University Press.

Richards, F.S.G. (1961). "A method of maximum likelihood estimation." J.R. Statist. Soc. B-23, 469-75.

Table 1

TOTAL MILES DRIVEN VS. FAMILY INCOME IN 1973

(adapted from Tables A6.1a and A6.1b of Holmes (1974) pp. 197-198)

Miles Driven (thousands)	Family Income (car owners)										Total
	0-3050	3050-4900	4900-6550	6550-8700	8700-10850	10850-12900	12900-15350	15350-18500	18500-23500	23500+	
0-5	123	141	99	89	82	41	46	18	18	6	663
5-10	42	91	109	134	102	69	68	44	51	39	749
10-15	31	50	102	109	130	146	116	100	76	61	921
15-20	15	16	32	46	60	75	80	97	68	58	547
20-25	7	18	20	44	51	53	51	60	62	60	426
25-30	2	6	7	12	19	30	42	33	44	37	232
30-35	1	6	10	16	18	25	23	28	19	43	189
35+	6	9	13	16	25	36	33	39	44	64	285
Total	227	337	392	466	487	475	459	419	382	368	4012

Figure 1

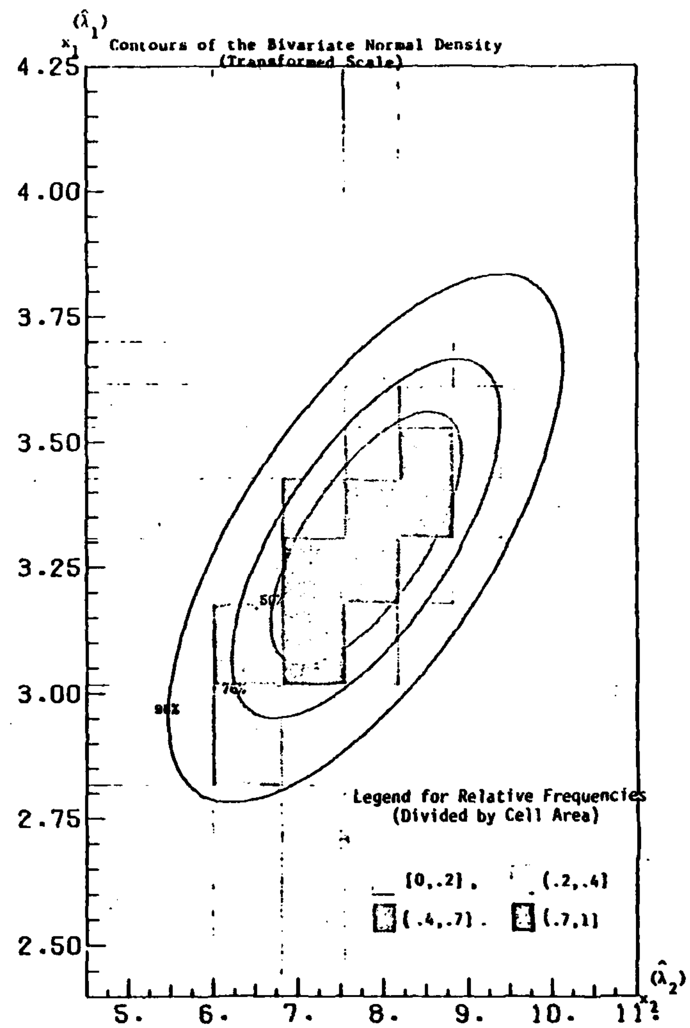


Figure 2

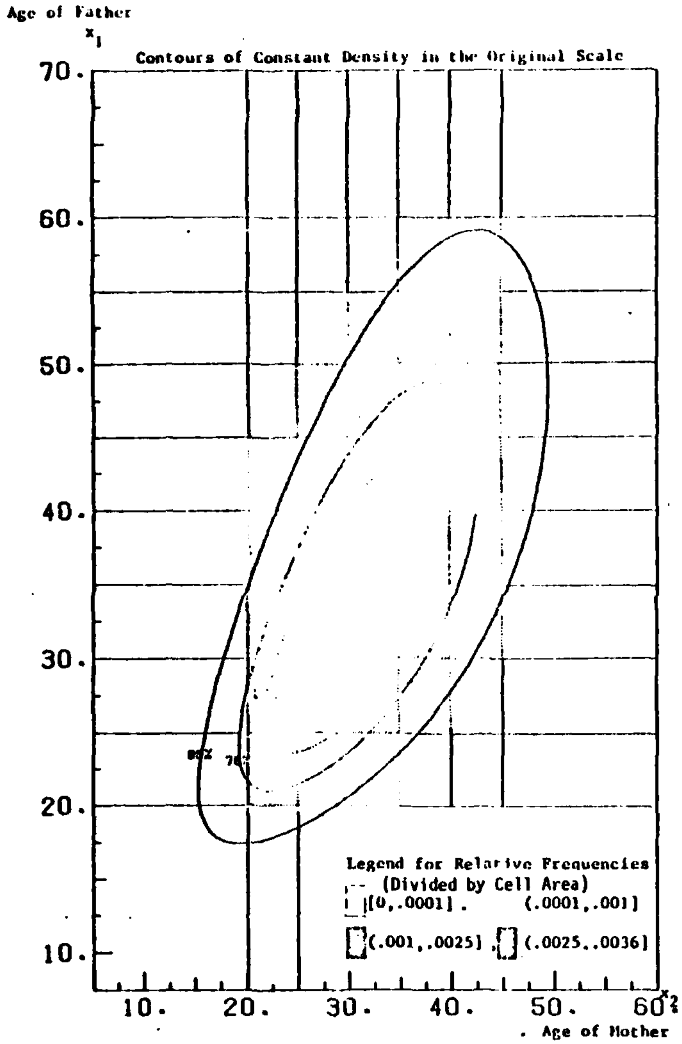


Figure 3

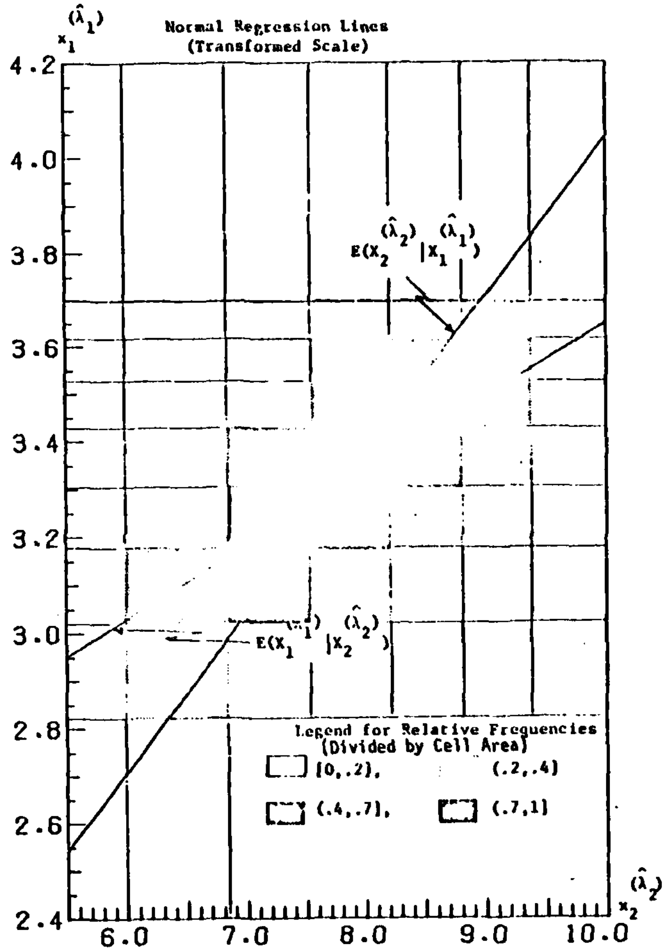


Figure 4

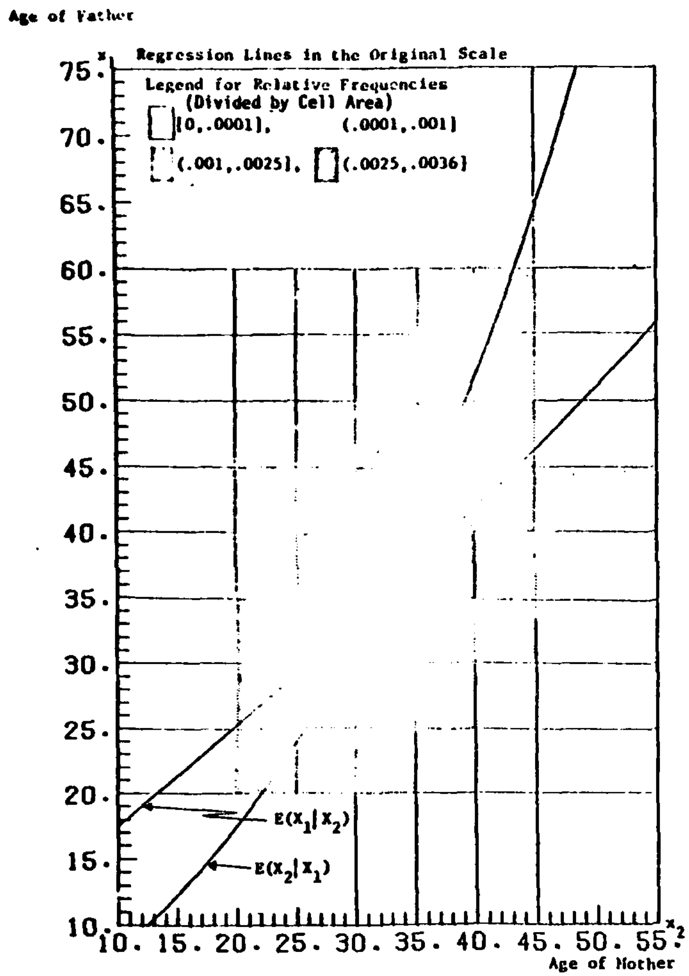


Figure 5

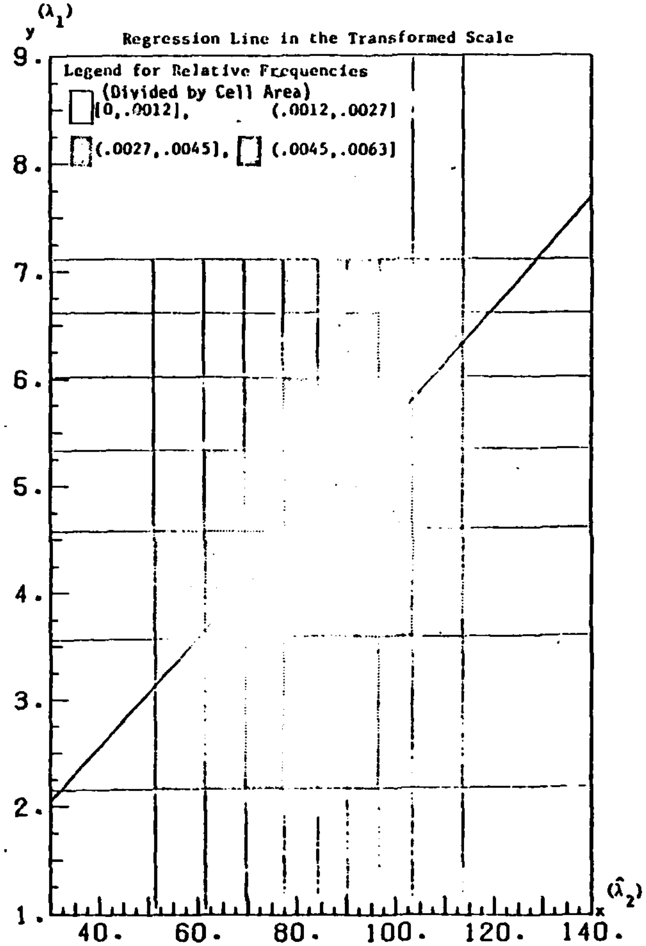
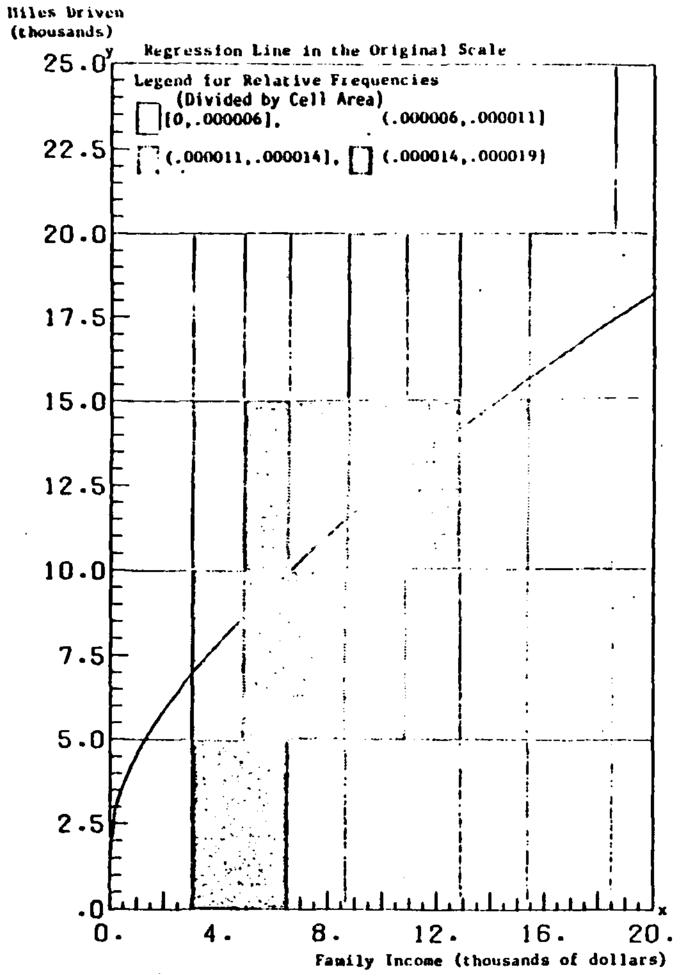


Figure 6



Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: Technical Report No. 592

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle): TRANSFORMING GROUPED BIVARIATE DATA TO NEAR NORMALITY

5. TYPE OF REPORT & PERIOD COVERED

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s): Victor M. Guerrero, Richard A. Johnson

8. CONTRACT OR GRANT NUMBER(s): ONR Grant No. N00014-78-C-07?? (Also funded by Army Res. Off.

9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Statistics, University of Wisconsin, Madison, Wisconsin 53706

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS: Office of Naval Research, 800 N. Quincy Street, Arlington, VA 22217

12. REPORT DATE: November 1979

13. NUMBER OF PAGES: 21 pages

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

15. SECURITY CLASS. (of this report): Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Distribution of this document is unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Grouped bivariate data

Transformations

Regression and correlation

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

We are concerned with the use of power transformations when data on two variables are presented in a two way table. Situations where one or both variables are grouped into intervals are considered and regressions and the correlation obtained from the transformed data. Also, by transforming back to the original scale, we obtain a smoothed version of the data.

DD FORM 1473 (1 JAN 73) EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)